Assignment 6: Web Crawler

Code:

```
val a = for {
    n <- ns \\ "a"
    hr <- n \ "@href"
} yield createURL(Option(url), hr.toString())
val b: Seq[URL] = a.flatMap(_.toOption)
for{
    x<-b
}yield validateURL(x)
```

```
for {
    a <- getURLContent(url)
    b <- MonadOps.asFuture(getLinks(a))
} yield b
```

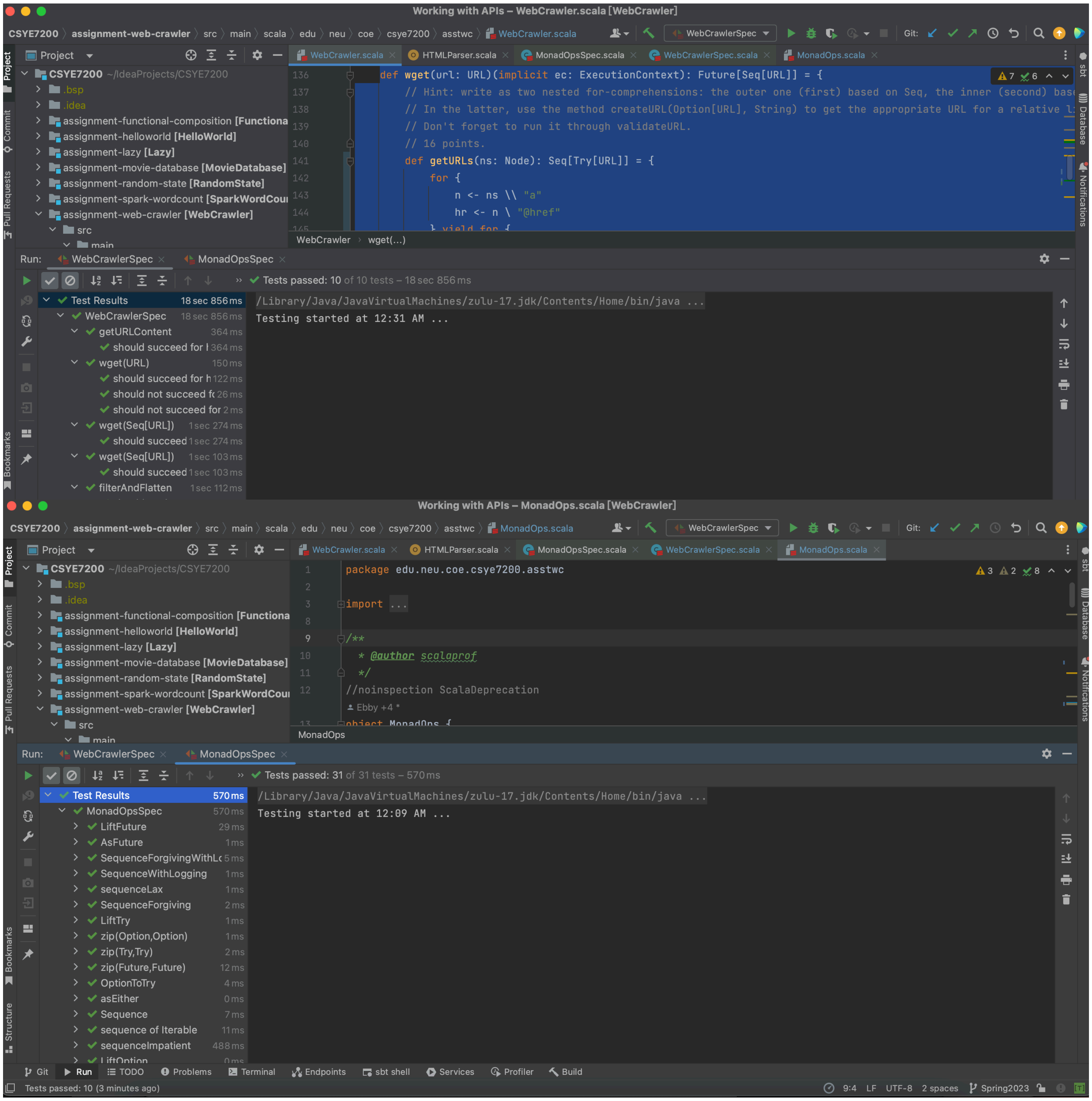
```
def asOption[X](xe: Either[Throwable, X]): Option[X]
=xe.toOption
```

Suggestions on how to improve the web crawler(get it from ChatGPT):

1. Separate concerns: It's a good practice to separate concerns in the code. Currently, the WebCrawler class is doing two things: performing the web crawl and logging errors. It would be better to separate the logging functionality into a separate class or method.
2. Error handling: The current error handling mechanism is not very robust. For example, if a URL is invalid or there is an error while reading the content of the URL, the error

is just logged to the console. It would be better to handle errors more gracefully, for example, by returning a future with a failure.

Screenshot:



Working with APIs – WebCrawler.scala [WebCrawler]
CSYE7200 〉 assignment-web-crawler 〉 src 〉 main 〉 scala 〉 edu 〉 neu 〉 coe 〉 csye7200 〉 asstwc 〉 WebCrawler.scala
WebCrawlerSpec
Git:
Project
CSYE7200 ~/IdeaProjects/CSYE7200
.bsp
.idea
assignment-functional-composition [Functiona
assignment-helloworld [HelloWorld]
assignment-lazy [Lazy]
assignment-movie-database [MovieDatabase]
assignment-random-state [RandomState]
assignment-spark-wordcount [SparkWordCou
assignment-web-crawler [WebCrawler]
src
main
WebCrawler.scala
HTMLParser.scala
MonadOpsSpec.scala
WebCrawlerSpec.scala
MonadOps.scala
WebCrawler › wget(...)
Run:
WebCrawlerSpec
MonadOpsSpec

```
def wget(url: URL)(implicit ec: ExecutionContext): Future[Seq[URL]] = {
  // Hint: write as two nested for-comprehensions: the outer one (first) based on Seq, the inner (second) base
  // In the latter, use the method createURL(Option[URL], String) to get the appropriate URL for a relative l
  // Don't forget to run it through validateURL.
  // 16 points.
  def getURLs(ns: Node): Seq[Try[URL]] = {
    for {
      n <- ns \\ "a"
      hr <- n \ "@href"
    } yield for {
```

Tests passed: 10 of 10 tests – 18 sec 856 ms

```
Test Results                      18 sec 856 ms
  WebCrawlerSpec                  18 sec 856 ms
    getURLContent                 364 ms
      should succeed for l        364 ms
    wget(URL)                     150 ms
      should succeed for h        122 ms
      should not succeed fc       26 ms
      should not succeed for      2 ms
    wget(Seq[URL])                1 sec 274 ms
      should succeed              1 sec 274 ms
    wget(Seq[URL])                1 sec 103 ms
      should succeed              1 sec 103 ms
    filterAndFlatten              1 sec 112 ms
```

```
/Library/Java/JavaVirtualMachines/zulu-17.jdk/Contents/Home/bin/java ...
Testing started at 12:31 AM ...
```

Working with APIs – MonadOps.scala [WebCrawler]
CSYE7200 〉 assignment-web-crawler 〉 src 〉 main 〉 scala 〉 edu 〉 neu 〉 coe 〉 csye7200 〉 asstwc 〉 MonadOps.scala
MonadOps

```
package edu.neu.coe.csye7200.asstwc

import ...

/**
  * @author scalaprof
  */
//noinspection ScalaDeprecation
Ebby +4 *
object MonadOps {
```

Tests passed: 31 of 31 tests – 570 ms

```
Test Results                      570 ms
  MonadOpsSpec                    570 ms
    LiftFuture                    29 ms
    AsFuture                      1 ms
    SequenceForgivingWithLc       5 ms
    SequenceWithLogging           1 ms
    sequenceLax                   1 ms
    SequenceForgiving             2 ms
    LiftTry                       1 ms
    zip(Option,Option)            1 ms
    zip(Try,Try)                  2 ms
    zip(Future,Future)            12 ms
    OptionToTry                   4 ms
    asEither                      0 ms
    Sequence                      7 ms
    sequence of Iterable          11 ms
    sequenceImpatient             488 ms
    LiftOption                    0 ms
```

```
/Library/Java/JavaVirtualMachines/zulu-17.jdk/Contents/Home/bin/java ...
Testing started at 12:09 AM ...
```

Tests passed: 10 (3 minutes ago)

Github: https://github.com/junlongqiao/CSYE7200